

Semiparametric Bayesian Information Criterion for Model Selection in Ultra-high Dimensional Additive Models

Heng Lian

Division of Mathematical Sciences

School of Physical and Mathematical Sciences

Nanyang Technological University

Singapore 637371

Singapore

July 26, 2011

Abstract

For linear models with a diverging number of parameters, it has recently been shown that modified versions of Bayesian information criterion (BIC) can identify the true model consistently. However, in many cases there is little justification that the effects of the covariates are actually linear. Thus a semi-parametric model such as the additive model studied here, is a viable alternative. We demonstrate that theoretical results on the consistency of BIC-type criterion can be extended to this more challenging situation, with dimension diverging exponentially fast with sample size. Besides, the noise assumptions

are relaxed in our theoretical studies. These efforts significantly enlarge the applicability of the criterion to a more general class of models.

Keywords: Bayesian information criterion (BIC); Selection consistency; Sparsity; Ultra-high dimensional models; Variable selection.

1 Introduction

With rapid increases in the production of large dimensional data by modern technology, more and more studies have focused on variable selection problems where the goal is to identify the few relevant predictors among a large collection of predictors, which might even outnumber the sample size due to the constraint of experimental costs. For example, in microarray experiments investigating genetic mechanisms of a certain disease, thousands of genes are assayed all at once while the number of samples is constrained by the cost of arrays as well as by the rarity of the disease in the population.

In linear models with fixed dimension, performance of various criteria for variable selection is well known (Shao, 1997), including AIC (Akaike, 1970), BIC (Schwarz, 1965), C_p (Mallows, 1973) etc. In particular, BIC was shown to be consistent in variable selection. More recently, penalization approaches to variable selection have drawn increasing attention due to their stability and computational attractiveness (Tibshirani, 1996; Yuan and Lin, 2006; Fan and Li, 2001; Zou, 2006; Wang et al., 2011). Following this trend, Wang et al. (2007) has shown that BIC computed along the solution path of the penalized estimator is also selection consistent.

Nevertheless, these traditional criteria are too liberal for regression problems with high dimensional covariates, in that they tend to incorporate many spurious covariates in the model selected. On the positive side, modifications of BIC by using a statistically motivated larger penalty term can successfully address this problem,

make the criterion provably consistent, and exhibit satisfactory performance in real applications (Wang et al., 2009; Chen and Chen, 2008). Despite these efforts, the works mentioned above, particularly the theoretical investigations, entirely focused on parametric linear models with Gaussian noise, while in many applications there is little a priori justification that the covariates actually have such simple linear effects on the responses.

The additive model introduced by Stone (1985) represents a more flexible class of semiparametric models that allows a general transformation of each covariate to enter as an additive component. This raises an interesting question: is there an appropriately modified BIC-type criterion that can consistently identify the nonzero components in this class of semiparametric models? Although a similar question has been answered in an affirmative way in Wang and Xia (2009) for fixed-dimensional varying-coefficient models, it remains a conjecture for high dimensional semiparametric problems. We note that Huang et al. (2010) has used modified BIC-type criterion in selecting the tuning parameter in group LASSO penalty for additive models, but they did not demonstrate the theoretical property of such a criterion. Compared to parametric models, the approximation errors for the component functions poses additional challenges to our analysis.

In this paper, we will investigate the theoretical property of BIC-type criterion in additive models with the number of components p growing much faster than sample size n . To be more specific, we assume $\log p = o(n^{2d/(2d+1)})$ where d characterizes the smoothness (roughly the number of derivatives) of the component functions. Following the existing literature, we say the problem has a ultra-high dimensionality. On the other hand, the number of truly nonzero components is assumed to be fixed and does not diverge with sample size, for the same reason as discussed in Huang et al. (2010). Besides, although we acknowledge that it might be restrictive to assume that all components have the same smoothness, it would be hard, if not impossible,

to satisfactorily deal with the more general case. Finally, it is worth noting that we relax the Gaussian noise assumption used in Chen and Chen (2008); Wang et al. (2009) to sub-Gaussian noise. The Gaussian assumption was key to make the theoretical analysis tractable in those studies (see for example (B.3) in Wang et al. (2009)). With sub-Gaussian noise, we need to resort to studying the tail probability of some quadratic forms involving sub-Gaussian random variables.

2 Bayesian Information Criterion for Unpenalized Polynomial Spline Estimators

Consider regression problems with observations $(Y_i, X_i), i = 1, \dots, n$ that are independent and identically distributed (i.i.d.) as (Y, X) , where Y is a scalar response and $X = (X_1, \dots, X_p)^T$ contains p covariates. Substantial progress has been made on linear regression when p is large, with or without penalty. Since fitting fully nonparametric models is infeasible for large dimensions, an elegant solution to relax the strong linearity assumption, known as the additive model (Stone, 1985; Hastie and Tibshirani, 1990), was proposed to avoid this difficulty, which is specified by

$$Y_i = \mu + \sum_{j=1}^p f_j(X_{ij}) + \epsilon_i, \tag{1} \quad \{\text{eqn:am}\}$$

where μ is the intercept, f_j are unknown univariate component functions and ϵ_i are i.i.d. mean zero noises.

Without loss of generality, we assume the distribution of X_j is supported on $[0, 1]$ and also impose the condition $Ef_j(X_j) = 0$ for identifiability. We use polynomial splines to approximate the components. Let $\tau_0 = 0 < \tau_1 < \dots < \tau_{K'} < 1 = \tau_{K'+1}$ be a partition of $[0, 1]$ into subintervals $[\tau_k, \tau_{k+1}), k = 0, \dots, K'$ with K' internal knots.

We only restrict our attention to equally spaced knots although data-driven choice can be considered such as putting knots at certain sample quantiles of the observed covariate values. A polynomial spline of order q is a function whose restriction to each subinterval is a polynomial of degree $q - 1$ and globally $q - 2$ times continuously differentiable on $[0, 1]$. The collection of splines with a fixed sequence of knots has a normalized B-spline basis $\{B_1(x), \dots, B_{\tilde{K}}(x)\}$ with $\tilde{K} = K' + q$. Because of the centering constraint $Ef_j(X_j) = 0$, we instead focus on the subspace of spline functions $S_j^0 := \{s : s = \sum_{k=1}^K b_{jk} B_{jk}(x), \sum_{i=1}^n s(X_{ij}) = 0\}$ with basis $\{B_{jk}(x) = B_k(x) - \sum_{i=1}^n B_k(X_{ij})/n, k = 1, \dots, K = \tilde{K} - 1\}$ (the subspace is $K = \tilde{K} - 1$ dimensional due to the empirical version of the constraint). Using spline expansions, we can approximate the components by $f_j(x) \approx \sum_k b_{jk} B_{jk}(x)$. Note that it is possible to specify different K for each component but we assume they are the same for simplicity (using the same K 's is reasonable when all components have the same smoothness parameter).

Suppose the true components are $f_{0j}, 1 \leq j \leq p$, and the true intercept is denoted by μ_0 . We consider a sparse model where only the first s components are nonzero. In unpenalized estimation, the following least squares estimation procedure is used to find the spline coefficients:

$$(\hat{\mu}, \hat{b}) = \arg \min_{\mu, b} \sum_i (Y_i - \mu - \sum_{j=1}^p \sum_{k=1}^K b_{jk} B_{jk}(X_{ij}))^2. \quad (2) \quad \{\text{eqn:min}\}$$

However, the resulting estimator cannot be consistent when p diverges at a sufficiently fast rate. Thus, we restrict our search on submodels where at most M components are nonzero, where M is a known fixed upper bound for s , and perform least squares regression with no more than M components in (2). Similar constraint is also imposed in Chen and Chen (2008) for linear models.

Let

$$Z_j = \begin{pmatrix} B_{j1}(X_{1j}) & B_{j2}(X_{1j}) & \cdots & B_{jK}(X_{1j}) \\ \vdots & \vdots & \vdots & \vdots \\ B_{j1}(X_{nj}) & B_{j2}(X_{nj}) & \cdots & B_{jK}(X_{nj}) \end{pmatrix}_{n \times K},$$

$Z = (Z_1, \dots, Z_p)$, $Y = (Y_1, \dots, Y_n)^T$. For any submodel indicated by $S \subseteq \{1, \dots, p\}$, let Z_S be the submatrix of Z containing the columns in S , and similarly defined b_S, \hat{b}_S , etc. For notation convenience, we add $(1, \dots, 1)/\sqrt{K}$ as the first column of Z, Z_S and define $a = (\sqrt{K}\mu, b^T)^T$, $a_S = (\sqrt{K}\mu, b_S^T)^T$, such that for the submodel S (2) can be written in matrix form as

$$\hat{a}_S = \min_{a_S} \|Y - Z_S a_S\|^2. \quad (3) \quad \{\text{eqn:min2}\}$$

Let the true model be indicated by $S_0 = \{1, \dots, s\}$.

Now we can define the BIC-type criterion for the semiparametric model as

$$BIC(S) = \log(\|Y - Z_S \hat{a}_S\|^2) + |S|K \frac{\log n + \log p}{n}, \quad (4) \quad \{\text{eqn:bic}\}$$

where $|S|$ is the size of the set S . The submodel \hat{S} that achieves the minimum value of the above (over all submodels with $|S| \leq M$) is chosen as the final model. The form of the above penalty is the same as that used in Huang et al. (2010) for group adaptive LASSO estimator, which is slightly different from that of Chen and Chen (2008), but easily seen to be asymptotically equivalent since $\log \binom{p}{j} \approx j \log p, j = 1, \dots, M$. The penalty in Wang et al. (2009), adapted to the semiparametric context here, is of the form $C_n |S| K \log n / n$ for some $C_n \rightarrow \infty$. We will try to be slightly more general and present our theoretical results for a general penalty term denoted by $pen(S)$.

The following technical conditions are assumed.

- (c1) The covariate vector X has a continuous density supported on $[0, 1]^p$. Furthermore, the marginal densities for $X_j, 1 \leq j \leq p$ are all bounded from below and

above by two fixed positive constants respectively.

- (c2) The mean zero noises ϵ_i are independent of covariates, have variance σ^2 , and are sub-Gaussian. That is there exists some $\alpha > 0$ such that $E[\exp\{t\epsilon\}] \leq \exp\{t^2\alpha^2/2\}$.
- (c3) $f_{0j}, 1 \leq j \leq s$ satisfies a Lipschitz condition of order $d > 1/2$: $|f_{0j}^{(\lfloor d \rfloor)}(t) - f_{0j}^{(\lfloor d \rfloor)}(s)| \leq C|s - t|^{d - \lfloor d \rfloor}$, where $\lfloor d \rfloor$ is the biggest integer strictly smaller than d and $f_{0j}^{(\lfloor d \rfloor)}$ is the $\lfloor d \rfloor$ -th derivative of f_{0j} . The order of the B-spline used satisfies $q \geq d + 2$.
- (c4) The number of nonzero components is $s = O(1)$.
- (c5) $K \log(pn)/n \rightarrow 0, K \rightarrow \infty, K \log(pn)/n + K^{-2d} = o(\min_{1 \leq j \leq s} \|f_{0j}\|^2), \text{pen}(S_0) = o(\min_{1 \leq j \leq s} \|f_{0j}\|^2), K^{-2d} = o(\text{pen}(S) - \text{pen}(S_0))$ for $S \supsetneq S_0, K \log(pn)/n = O(\text{pen}(S) - \text{pen}(S_0))$ for $S \supsetneq S_0$.

Most of the assumptions are standard in the literature. Assumptions (c1)-(c4) are also assumed in Huang et al. (2010). However, we will not assume that $\min_{1 \leq j \leq s} \|f_{0j}\|$ is bounded away from zero as in assumption (A1) of Huang et al. (2010). Instead, (c5) makes it clear that this quantity is allowed to converge to zero at a certain rate. Also note that in previous studies on the consistency of BIC-type criterion in linear models, Gaussian noise is assumed. We relax this assumption at the cost of more sophisticated arguments. We collect the assumptions on the convergence/divergence rate of different quantities in (c5). The expressions in (c5) can be simplified when $K \sim n^{1/(2d+1)}$ (this is the theoretically optimal choice of K that balances bias and variance (Stone, 1985)) and $\text{pen}(S) = |S|K(\log n + \log p)/n$ (see Corollary 1 below).

Theorem 1 *Assume conditions (c1)-(c5). Then*

$$P(\hat{S} = S_0) \rightarrow 1.$$

By this theorem, we know that with probability tending to 1, any model with size no larger than M cannot be selected by BIC-type criterion, other than the true one. For particular form of the penalty function stated above, we have the following corollary.

Corollary 1 *If $K \sim n^{1/(2d+1)}$, $\log p = o(n^{2d/(2d+1)})$, $\min_{1 \leq j \leq s} \|f_{0j}\|^2 \gg (\log(pn))n^{-2d/(2d+1)}$, then under conditions (c1)-(c4) the BIC-type criterion defined in (4) is selection consistent.*

3 Bayesian Information Criterion for Penalized Estimators

In the last section we stated that BIC-type criterion is consistent for variable selection for unpenalized estimators. However, even when the size of the submodels under consideration is constrained by M , brute-force search is still infeasible for large p . This is one of the reasons why penalized estimators become so popular in recent years. Here we briefly discuss how the results in the previous section can be extended to penalized estimator.

In our context, the penalized estimator is defined by

$$\hat{a}_\lambda = \arg \min \|Y - Za\|^2 + \sum_{j=1}^p p_\lambda(\|b_j\|), \quad (5) \quad \{\text{eqn:pen}\}$$

where λ is the tuning parameter controlling the sparsity of the solution, with larger λ resulting in more components estimated as zero. Let $S_\lambda = \{j : \hat{b}_{\lambda j} \neq 0\}$ be the submodel represented by \hat{a}_λ . Here we focus on the group adaptive LASSO penalty since this is the one studied in Huang et al. (2010) for ultra-high dimensional additive models, although we expect selection consistency for estimators with SCAD penalty

(Fan and Li, 2001) or MCP (Zhang, 2010) can be derived in a similar way. Thus we assume all the conditions in Huang et al. (2010). The BIC-type criterion for penalized estimator is defined as

$$BIC(\lambda) = \log(\|Y - Z\hat{a}_\lambda\|^2) + \text{pen}(S_\lambda),$$

and the optimal tuning parameter is $\hat{\lambda} = \arg \min_{\lambda > 0} BIC(\lambda)$.

Following Huang et al. (2010), for the group adaptive LASSO estimator, the penalty term in (5) is $\sum_{j=1}^p \lambda \|b_j\| / \|\tilde{b}_j\|$ where $\|\tilde{b}_j\|$ is the initial group LASSO estimator. The following discussions are mainly extensions of arguments in Wang et al. (2009). Based on Corollary 2 in Huang et al. (2010), if $K \sim n^{1/(2d+1)}$ and the tuning parameter is chose to be $\lambda_n \sim \sqrt{n}$, the estimator \hat{a}_{λ_n} represents the correct model (that is $\hat{b}_{\lambda_n j} = 0$ for $j > s$, or in other words $S_{\lambda_n} = S_0$). Since $\hat{b}_{\lambda_n j} = 0$ for $j > s$, $\hat{a}_{\lambda_n S_0} = (\sqrt{K}\hat{\mu}_{\lambda_n}, \hat{b}_{\lambda_n 1}, \dots, \hat{b}_{\lambda_n s})^T$ must be the minimizer of

$$\|Y - Z_{S_0}a\|^2 + \sum_{j=1}^s \lambda_n \|b_j\| / \|\tilde{b}_j\|,$$

which yields by first order condition $\hat{a}_{\lambda_n S_0} = (Z_{S_0}^T Z_{S_0})^{-1}(Z_{S_0}^T Y + \nu)$, where

$$\nu = \partial \sum_{j=1}^s p_\lambda(\|b_j\|) / \partial a \Big|_{a=\hat{a}_{\lambda_n S_0}} = \lambda_n (0, \frac{\hat{a}_{\lambda_n 1}^T}{\|\tilde{a}_1\| \cdot \|\hat{a}_{\lambda_n 1}\|}, \dots, \frac{\hat{a}_{\lambda_n s}^T}{\|\tilde{a}_s\| \cdot \|\hat{a}_{\lambda_n s}\|})^T.$$

We have $\|\nu\|^2 = O(\lambda_n^2/K) = O(n/K)$. Thus

$$\begin{aligned}
& \|Y - Z\hat{a}_{\lambda_n}\|^2 - \|Y - Z_{S_{\lambda_n}}\hat{a}_{S_{\lambda_n}}\|^2 \\
&= \|Z_{S_0}(Z_{S_0}^T Z_{S_0})^{-1}\nu\|^2 - 2(Y - P_{S_0}Y)(Z_{S_0}Z_{S_0}^T Z_{S_0})^{-1}\nu \\
&= O((K/n)\|\nu\|^2 + \sqrt{K + n/K^{2d}}\sqrt{K/n}\|\nu\|) \\
&= O(\sqrt{K + n/K^{2d}}).
\end{aligned}$$

Thus

$$\begin{aligned}
& BIC(\lambda) - BIC(\lambda_n) \\
&= \log(\|Y - Z\hat{a}_\lambda\|^2) - \log(\|Y - Z\hat{a}_{\lambda_n}\|^2) + \text{pen}(S_\lambda) - \text{pen}(S_{\lambda_n}) \\
&\geq \log(\|Y - Z\hat{a}_{S_\lambda}\|^2) - \log(\|Y - Z\hat{a}_{\lambda_n}\|^2) + \text{pen}(S_\lambda) - \text{pen}(S_{\lambda_n}) \\
&= \log(\|Y - Z\hat{a}_{S_\lambda}\|^2) - \log(\|Y - Z\hat{a}_{S_{\lambda_n}}\|^2) + \text{pen}(S_\lambda) - \text{pen}(S_{\lambda_n}) + O(\sqrt{K + n/K^{2d}}) \\
&= BIC(S_\lambda) - BIC(S_0) + O(\sqrt{K + n/K^{2d}}).
\end{aligned}$$

A look at the proof for Theorem 1 in the Appendix shows that when $S_\lambda \neq S_0$ the gap between $BIC(S_\lambda)$ and $BIC(S_0)$ is actually larger than $O(\sqrt{K + n/K^{2d}})$, so the $O(\sqrt{K + n/K^{2d}})$ actually does not affect the result and we still have $BIC(\lambda) - BIC(\lambda_n) > 0$ with probability tending to 1 uniformly over all λ such that $S_\lambda \neq S_0$ and $|S_\lambda| \leq M$.

4 Conclusion and Discussion

In this paper, we showed that the BIC-type criterion can be used in additive models with ultra-high feature dimensions to consistently select the true model. This paper is mainly of theoretical interest, and numerical evidence of its performance was con-

tained already in Huang et al. (2010). Although the BIC-type criterion is consistent for both unpenalized and the penalized estimators, computational constraints imply that the latter should be used in practice to avoid brute-force search over submodels. When the dimension of the feature space is so high that penalized approaches cannot be directly applied due to computational reasons, nonparametric independence screening procedure (Fan et al., 2011) can be used as a first step to reduce the dimensionality.

The BIC-type criterion for penalized estimator focuses on the choice of tuning parameter λ and ignores the choice of K (the number of knots in B-spline approximation). In practice, K can be fixed to a reasonable integral value as done in Yu and Ruppert (2002); Huang et al. (2010); Fan et al. (2011) and some sensitivity analysis might be justified. It remains an open problem whether some criterion exists for data-driven choice of K in high-dimensional contexts that has the desired theoretical property (in particular results in $K \sim n^{1/(2d+1)}$).

Appendix: Proofs

By well-known properties of B-splines, there exists $b_{0j} = (b_{0j1}, \dots, b_{0jK})^T$ that satisfies the approximation property $\|\sum_k b_{0jk} B_{jk}(x) - f_{0j}(x)\|_\infty = O(K^{-d})$. Let $a_0 = (\sqrt{K}\mu_0, b_{01}^T, \dots, b_{0p}^T)^T$ and similarly define a_{0S} for a submodel S . In our proofs, C denotes a generic positive constant. We first present a Lemma which will be useful in the proof of the Theorem.

Lemma 1

$$\sup_{S \supseteq S_0: |S| \leq M} \left| \|Y - Z_S \hat{a}_S\|^2 - \|Y - Z_S a_{0S}\|^2 \right| = O(nK^{-2d}) + o(K \log(pn)).$$

Proof of Lemma 1. We have

$$\begin{aligned}
& \|Y - Z_S \hat{a}_S\|^2 - \|Y - Z_S a_{0S}\|^2 \\
&= -2(Y - Z_S a_{0S})^T Z_S (\hat{a}_S - a_{0S}) + \|Z_S (\hat{a}_S - a_{0S})\|^2 \\
&= -2\epsilon^T Z_S (\hat{a}_S - a_{0S}) - 2(f_0(X) - Z_S a_{0S})^T Z_S (\hat{a}_S - a_{0S}) + \|Z_S (\hat{a}_S - a_{0S})\|^2,
\end{aligned} \tag{6}$$

where $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$ and $f_0(X) = (f_0(X_1), \dots, f_0(X_n))^T$ with $f_0(X_i) = \mu_0 + \sum_{j=1}^s f_{0j}(X_{ij})$ being the true regression function evaluation at covariate X_i .

By definition we have $\hat{a}_S = (Z_S^T Z_S)^{-1} Z_S^T (Z_S^T a_{0S} + (f_0(X) - Z_S^T a_{0S}) + \epsilon)$ and thus $\hat{a}_S - a_{0S} = (Z_S^T Z_S)^{-1} Z_S^T (f_0(X) - Z_S^T a_{0S}) + (Z_S^T Z_S)^{-1} Z_S^T \epsilon$. Plugging this expression into (6) we get

$$\begin{aligned}
& \|Y - Z_S \hat{a}_S\|^2 - \|Y - Z_S a_{0S}\|^2 \\
&= -2\epsilon^T P_S \epsilon - 4\epsilon^T P_S (f_0(X) - Z_S^T a_{0S}) \\
&\quad - 2(f_0(X) - Z_S^T a_{0S})^T P_S (f_0(X) - Z_S^T a_{0S}) + \|P_S \epsilon + P_S (f_0(X) - Z_S^T a_{0S})\|^2 \\
&= O(\epsilon^T P_S \epsilon + (f_0(X) - Z_S^T a_{0S})^T P_S (f_0(X) - Z_S^T a_{0S})),
\end{aligned} \tag{7}$$

where $P_S = Z_S (Z_S^T Z_S)^{-1} Z_S^T$ is a projection matrix.

Obviously $(f_0(X) - Z_S^T a_{0S})^T P_S (f_0(X) - Z_S^T a_{0S}) = O(nK^{-2d})$. Next we will show $\sup_{S: |S| \leq M} \epsilon^T P_S \epsilon = o(K \log(pn))$. Since we do not assume the errors are Gaussian, the quadratic form cannot be written as sum of chi-squared random variables. Fortunately we can still resort to results on quadratic forms for sub-Gaussian random variables. Specifically by Proposition 1.1 in Mikosch (1991), when $y \geq K\alpha^2$, we have

$$P(\epsilon^T P_S \epsilon > \alpha^2 M K + y) \leq \exp\{-Cy/\alpha^2\},$$

and thus

$$P\left(\sup_{S:|S|\leq M} \epsilon^T P_S \epsilon > \alpha^2 M K + y\right) \leq O(p^M) \exp\{-Cy/\alpha^2\},$$

and if one takes $y = \delta K \log(pn)$ for any $\delta > 0$, the above probability will tend to 0. This shows $\sup_{S:|S|\leq M} \epsilon^T P_S \epsilon = o(K \log(pn))$. \square

Proof of Theorem 1. The proof is split into two parts, considering the underfitted models (some nonzero components are not in S) and overfitted models (some zero components, as well as all nonzero components, are included in S) respectively.

Part 1: $S_0 \not\subseteq S$.

Let \hat{a}_S and \hat{a}_{S_0} be the least squares estimator under submodel S and the true model S_0 respectively. Let $\tilde{S} = S \cup S_0$. With abuse of notation, \hat{a}_S is also used to denote $|\tilde{S}|K+1$ -dimensional vector where the coefficients not associated the submodel S is filled in by zero. Similar statement applies to other notations such as a_{S_0} , \hat{a}_{S_0} etc. Thus we can write expressions such as $Z_{\tilde{S}}\hat{a}_S$ even though $\tilde{S} \neq S$. That is, zero values are filled in to match the dimension. Then we have

$$\begin{aligned} & \|Y - Z_{\tilde{S}}\hat{a}_S\|^2 - \|Y - Z_{\tilde{S}}\hat{a}_{S_0}\|^2 \\ &= -2(Y - Z_{\tilde{S}}\hat{a}_{S_0})^T Z_{\tilde{S}}(\hat{a}_S - \hat{a}_{S_0}) + \|Z_{\tilde{S}}(\hat{a}_S - \hat{a}_{S_0})\|^2 \\ &= -2\epsilon^T Z_{\tilde{S}}(\hat{a}_S - \hat{a}_{S_0}) + 2(Z_{\tilde{S}}^T \hat{a}_{S_0} - f_0(X))^T Z_{\tilde{S}}(\hat{a}_S - \hat{a}_{S_0}) + \|Z_{\tilde{S}}(\hat{a}_S - \hat{a}_{S_0})\|^2. \end{aligned} \tag{8}$$

By existing results on spline estimator in additive models (Stone, 1985), we know that when the true model is known, $\|\hat{a}_{S_0} - a_{0S_0}\| = O(K/\sqrt{n} + K^{-d+1/2})$. Besides, since some nonzero components in a_{S_0} is estimated as zero in \hat{a}_S , we know $\|\hat{a}_S - a_{0S_0}\| \geq \min_{1 \leq j \leq s} \|b_{0j}\| \geq C\sqrt{K}(\min_{1 \leq j \leq s} \|f_{0j}\| - K^{-d})$ by the approximate property of splines.

Thus uniformly for all $S \not\supseteq S_0$,

$$\|\hat{a}_S - \hat{a}_{S_0}\| \geq \|\hat{a}_S - a_{0S_0}\| - \|a_{0S_0} - \hat{a}_{S_0}\| \geq C(\sqrt{K} \min_{1 \leq j \leq s} \|f_{0j}\| - K/\sqrt{n} - K^{-d+1/2}).$$

Denote the right hand side above by γ_n , then the third term in (8) is bounded below by $C(n/K)\gamma_n^2$ by Lemma 3 in Huang et al. (2010). The absolute value of the second term is bounded by $\sqrt{nK^{-2d}}\sqrt{n/K}\gamma_n$ and thus is of smaller order than the third term. Finally we bound the first term in (8) by

$$-2\epsilon^T Z_{\hat{S}}(\hat{a}_S - \hat{a}_{S_0}) \geq -4\epsilon^T P_{\hat{S}}\epsilon - \frac{1}{4}\|Z_{\hat{S}}(\hat{a}_S - \hat{a}_{S_0})\|^2.$$

In the proof of Lemma 1 we showed that $\sup_{S:|S|\leq M} \epsilon^T P_{\hat{S}}\epsilon = o(K \log(pn))$ and thus by condition (c5), (8) is bounded below a positive number at least as large as $C(n/K)\gamma_n^2$. We have

$$\begin{aligned} & BIC(S) - BIC(S_0) \\ = & \log \left(1 + \frac{\|Y - Z_S \hat{a}_S\|^2/n - \|Y - Z_{S_0} \hat{a}_{S_0}\|^2/n}{\|Y - Z_{S_0} \hat{a}_{S_0}\|^2/n} \right) + pen(S) - pen(S_0). \end{aligned}$$

Lemma 1 implies that $\|Y - Z_{S_0} \hat{a}_{S_0}\|^2/n \geq \|Y - Z_{S_0} a_{0S_0}\|^2/n - O(K^{-2d}) - o((K/n) \log(pn)) \geq \|\epsilon\|^2/(2n) - \|Z_{S_0} a_{0S_0} - f_0(X)\|^2/n - O(K^{-2d}) - o((K/n) \log(pn)) \rightarrow \sigma^2/2$. Thus

$$\begin{aligned} & BIC(S) - BIC(S_0) \\ \geq & C(\min_{1 \leq j \leq s} \|f_{0j}\|^2 - K/n - K^{-2d}) + pen(S) - pen(S_0), \end{aligned}$$

which is positive with probability tending to 1 by (c5). Thus $P(\min_{S \not\supseteq S_0: |S| \leq M} BIC(S) - BIC(S_0) > 0) \rightarrow 1$.

Part 2: $S \supsetneq S_0$.

Lemma 1 showed that

$$\sup_{S \supseteq S_0} \|Y - Z_S a_{0S}\|^2 - \|Y - Z_S \hat{a}_S\|^2 = O(nK^{-2d}) + o(K \log(pn)), \quad (9) \quad \{\text{eqn:lem}\}$$

and noting that $Z_S a_{0S} = Z_{S_0} a_{0S_0}$ for $S_0 \subseteq S$, we have

$$\begin{aligned} & BIC(S_0) - BIC(S) \\ = & \log \left(\frac{\|Y - Z_{S_0} \hat{a}_{S_0}\|^2}{\|Y - Z_S \hat{a}_S\|^2} \right) - (pen(S) - pen(S_0)) \\ \leq & \log \left(\frac{\|Y - Z_{S_0} a_{0S_0}\|^2}{\|Y - Z_S \hat{a}_S\|^2} \right) - (pen(S) - pen(S_0)) \\ = & \log \left(1 + \frac{\|Y - Z_S a_{0S}\|^2 - \|Y - Z_S \hat{a}_S\|^2}{\|Y - Z_S \hat{a}_S\|^2} \right) - (pen(S) - pen(S_0)). \end{aligned}$$

Using (9) and similar to the arguments at the end of Part 1, $\|Y - Z_S \hat{a}_S\|^2/n$ is bounded away from zero uniformly in $S \supseteq S_0$. And thus $BIC(S_0) - BIC(S) \leq O(K^{-2d}) + o(K \log(pn)/n) - (pen(S) - pen(S_0)) < 0$ with probability tending to 1. \square

References

- Akaike, H. “Statistical predictor identification.” *Annals of the Institute of Statistical Mathematics*, 22:203–217 (1970).
- Chen, J. and Chen, Z. “Extended Bayesian information criteria for model selection with large model spaces.” *Biometrika*, 95(3):759–771 (2008).
- Fan, J., Feng, Y., and Song, R. “Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models.” *Journal of the American Statistical Association*, 106:544–557 (2011).

- Fan, J. Q. and Li, R. Z. “Variable selection via nonconcave penalized likelihood and its oracle properties.” *Journal of the American Statistical Association*, 96(456):1348–1360 (2001).
- Hastie, T. and Tibshirani, R. *Generalized additive models*. Monographs on statistics and applied probability. London ; New York: Chapman and Hall, 1st edition (1990).
- Huang, J., Horowitz, J. L., and Wei, F. “Variable selection in nonparametric additive models.” *Annals of Statistics*, 38(4):2282–2313 (2010).
- Mallows, C. “Some comments on C_p .” *Technometrics*, 15:661–675 (1973).
- Mikosch, T. “Estimates for tail probabilities of quadratic and bilinear forms in sub-gaussian random variables with applications to the low iterated logarithm.” *Probability and Mathematical Statistics*, 11:169–178 (1991).
- Schwarz, L. “On Bayes Procedures.” *Z. Wahrsch. Verw. Gebiete*, 4:10–26 (1965).
- Shao, J. “An asymptotic theory for linear model selection.” *Statistica Sinica*, 7(2):221–242 (1997).
- Stone, C. “Additive regression and other nonparametric models.” *The annals of Statistics*, 13:689–705 (1985).
- Tibshirani, R. “Regression shrinkage and selection via the Lasso.” *Journal of the Royal Statistical Society Series B-Methodological*, 58(1):267–288 (1996).
- Wang, H., Li, R., and Tsai, C. L. “Tuning parameter selectors for the smoothly clipped absolute deviation method.” *Biometrika*, 94(3):553–568 (2007).
- Wang, H. S., Li, B., and Leng, C. L. “Shrinkage tuning parameter selection with a diverging number of parameters.” *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 71:671–683 (2009).

- Wang, H. S. and Xia, Y. C. “Shrinkage estimation of the varying coefficient model.” *Journal of the American Statistical Association*, 104(486):747–757 (2009).
- Wang, L., Liu, X., Liang, H., and Carroll, R. “Estimation and variable selection for generalized additive partial linear models.” *Annals of Statistics*, to appear (2011).
- Yu, Y. and Ruppert, D. “Penalized spline estimation for partially linear single-index models.” *Journal of the American Statistical Association*, 97(460):1042–1054 (2002).
- Yuan, M. and Lin, Y. “Model selection and estimation in regression with grouped variables.” *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 68:49–67 (2006).
- Zhang, C. “Nearly unbiased variable selection under minimax concave penalty.” *The Annals of Statistics*, 38(2):894–942 (2010).
- Zou, H. “The adaptive lasso and its oracle properties.” *Journal of the American Statistical Association*, 101(476):1418–1429 (2006).